Data Research Engineer Assignment

There are two distinct sections to this assignment. The first is a data analysis task with a well-defined goal. The second is a more open-ended data exploration task, to help us understand how you approach working with messy, unstructured data. We expect the first task to take no more than 1 hour. The second task is less clear cut, however we suggest that you restrict yourself to a maximum of 2 hours — there is no perfect solution, and our priority with this task is to understand the steps you might take when faced with a new dataset and a vaguely-defined goal. It's more about the process than a specific (or complete) solution.

Instructions

- Please use Python and complete each task in a separate Colab/Jupyter notebook.
- The tasks are designed to show us how you think through a data problem, so please show all your workings and any analysis that leads you in a particular direction. Don't worry if it's messy — we're not expecting a perfect report. We will spend some time in the follow-up interview discussing your work.
- Please document your thought process with comments/text as you go along.
- Make it clear where you have made any assumptions about the data or problem.

Task 1 (data_task_1.csv)

This dataset has six columns: *Input1*, *Input2*, *Input3*, *Input4*, *Input5*, and *Output*. There are 10,000 rows of numeric data. Please investigate the data, and note down what you can about the dataset, showing whatever analyses you feel are appropriate. Ultimately, "Output" can be predicted quite accurately using the five "Input" columns and a simple mathematical formula. Please aim to find this formula. Whether you find it or not, we are interested in your analyses and thought processes to understand the data.

Task 2 (data_task_2.csv)

The motivation behind this task is the need for clean, high-quality datasets on which to train large machine learning models. The task itself is about producing a high-quality English-language news summarisation dataset, to be used for training a text summarisation model.

This dataset contains three columns: *url*, *scraped_article*, and *summary*. It has 40,000 rows of text data. The three columns for each data point are defined as follows:
***url*:** The URL of an online news article.
scraped_article: The text obtained from scraping the web page at the given URL, using a generic news article web scraper [note: no specific web scraping knowledge is required].
summary: A summary of the article, written by a human, based on the original web page (i.e. not based on the scraped text).

Your task is to explore the data and outline the steps you would take to pre-process the data in order to improve the quality of the dataset for model training. Note that removal of data points is perfectly acceptable — we don't expect the processed dataset to be the same size as the original (assume for the purpose of this task that we are not short of data).

We *don't* expect you to create a perfect finished product. Mainly, we would like to see what steps you take to decide how to process the data. It would be useful to see code sketches for any transformation or filtering steps you propose, but there's no need to turn these into robust implementations. Please also include any analysis you do to inform your decisions.